Anal. Chem. **2001**, *73*, 4566−4573

Bioinformatics and Mass Spectrometry for Microorganism Identification: Proteome-Wide Post-Translational Modifications and Database Search Algorithms for Characterization of Intact *H. pylori*

Plamen A. Demirev,*,† Jeffrey S. Lin,‡ Fernando J. Pineda,‡ and Catherine Fenselau†

Department of Chemistry, University of Maryland, College Park, Maryland 20742, and Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland 20723

MALDI-TOF mass spectrometry has been coupled with Internet-based proteome database search algorithms in an approach for direct microorganism identification. This approach is applied here to characterize intact *H. pylori* (strain 26695) Gram-negative bacteria, the most ubiquitous human pathogen. A procedure for including a specific and common posttranslational modification, N-terminal Met cleavage, in the search algorithm is described. Accounting for posttranslational modifications in putative protein biomarkers improves the identification reliability by at least an order of magnitude. The influence of other factors, such as number of detected biomarker peaks, proteome size, spectral calibration, and mass accuracy, on the microorganism identification success rate is illustrated as well.

Rapid and reliable identification of microorganisms has become an analytical challenge of increasing importance to many constituencies, including those involved in food safety, medical diagnostics, and counterterrorism.[1] More than 25 years ago, mass spectrometry was identified as a potentially viable physical method for characterization of microbial samples on the basis of the detection of specific biomarker molecules.[2,3] The advent of newer ionization techniques in the past decade has advanced the prospects to develop robust, automated, and miniaturized mass spectrometry-based systems for applications in microbiology.[4] In particular, observation of unique protein biomarker patterns in MALDI-TOF mass spectra from lysed and intact microorganisms[5−13] has focused subsequent research[14,15] into several logical avenues. One approach is the generation and compilation of fingerprint libraries of mass spectra from a variety of microorganism sources,[16] which raises the related issues of standardization, reproducibility, and accuracy of mass spectral collection and data analysis.[9,17] In addition, phenotyping of pathogenic *Escherichia coli* strains has been performed by clustering MALDI mass spectra using simple distance-based criteria.[18] Sample preparation methodologies[19−23] have been underlined as an important factor in developing sensitive MS-based methods for microorganism identification. Research has commenced to elucidate the structures of observed individual biomarkers from intact bacterial cells,[24,25]

* To whom correspondence should be addressed. Phone: 301-405-8618. Fax: 301-405-8615. E-mail: demirev@wam.umd.edu.

† University of Maryland.

‡ Johns Hopkins University.

(1) Walt, D. R.; Franz, D. R. *Anal. Chem.* **2000**, *72*, 738A−746A.

(2) Anhalt, J.; Fenselau, C. *Anal. Chem.* **1975**, *47*, 219−225.

(3) Fenselau, C., Ed.; "*Mass Spectrometry for the Characterization of Microorganisms*", ACS Symposium Series; American Chemical Society: Washington, DC, **1994**, Vol. 541.

(4) Bryden, W.; Benson, R.; Ecelberger, S.; Philips, T.; Cotter, R.; Fenselau, C. *Johns Hopkins APL Techn. Digest* **1995**, *16*, 296−310.

(5) Cain, T.; Lubman, D. M.; Weber, W. J. *Rapid Commun. Mass Spectrom.* **1994**, *8*, 1026−1030.

(6) Claydon, M. A.; Davey, S. N.; Edwards-Jones, V.; Gordon, D. B. *Nat. Biotechnol.* **1996**, *14*, 1584−1586.

(7) Krishnamurthy, T.; Ross, P. L.; Rajamani, U. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 883−888.

(8) Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay, J. O., Jr. *Rapid Commun. Mass Spectrom.* **1996**, *10*, 1227−32.

(9) Wang, Z.; Russon, L.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1998**, *12*, 456−464.

(10) Arnold, R. J.; Karty, J. A.; Ellington, A. D.; Reilly, J. P. *Anal. Chem.* **1999**, *71*, 1990−6.

(11) Lynn, E.; Chung, M.; Tsai, W.; Han, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 2022−2027.

(12) Haag, A.; Taylor, S.; Johnston, K.; Cole, R. *J. Mass Spectrom.* **1998**, *33*, 750−756.

(13) Hathout, Y.; Demirev, P. A.; Ho, Y. P.; Bundy, J. L.; Ryzhov, V.; Sapp, L.; Stutler, J.; Jackman, J.; Fenselau, C. *Appl. Environ. Microbiol.* **1999**, *65*, 4313−4319.

(14) Lay, J. O. *TRAC-Trend Anal. Chem.* **2000**, *19*, 507−516.

(15) Van Baar, B. L. M. *FEMS Microbiol. Rev.* **2000**, *24*, 193−219.

(16) Jarman, K. H.; Cebula, S. T.; Saenz, A. J.; Petersen, C. E.; Valentine, N. B.; Kingsley, M. T.; Wahl, K. L. *Anal. Chem.* **2000**, *72*, 1217−1223.

(17) Gantt, L.; Valentine, N.; Saenz, A.; Kingsley, M.; Wahl, K. L. *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 1131−1137.

(18) Conway, G.; Smole, S.; Sarracino, D.; Arbeit, R.; Leopold, P. *J. Microbiol. Mol. Biotechnol.* **2001**, *3*, 103−112.

(19) Birmingham, J.; Demirev, P.; Ho, Y. P.; Thomas, J.; Bryden, W.; Fenselau, C. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 604−607.

(20) Bundy, J. L.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 751−757.

(21) Kim, Y. J.; Freas, A.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 1544−1548.

(22) Madonna, A. J.; Basile, F.; Ferrer, I.; Meetani, M. A.; Rees, J. C.; Voorhees, K. J. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 2220−2229.

(23) Evason, D. J.; Claydon, M. A.; Gordon, D. B. *Rapid Commun. Mass Spectrom.* **2000**, *14*, 669−672.

(24) Holland, R. D.; Duffy, C. R.; Rafii, F.; Sutherland, J. B.; Heinze, T.; Holder, C. L.; Voorhees, K. J.; Lay, J. O. *Anal. Chem.* **1999**, *71*, 3226−3230.

(25) Dai, Y.; Li, L.; Roser, D. C.; Long, S. R. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 73−78.

10.1021/ac010466f CCC: $20.00
Published on Web 08/30/2001

spores,[26] and viruses,[21] and to correlate the appearance of biomarker peaks in bacterial mass spectra with a number of physical and chemical parameters of individual proteins.[27]

Recently, we developed bioinformatics tools to incorporate proteome database search algorithms for microorganism identification by mass spectrometry.[28,29] The approach is based on experimentally determining the masses, M_r, of a set of protein biomarkers from intact unknown organisms. Protein "hit lists" for different microorganisms are compiled by matching M_r against sequence-derived theoretical M_r of proteins (retrieved together with their organismic sources from Internet-accessible databases). A microorganism is ranked according to the number of matched mass peaks by combining the lists for all of the peaks. This straightforward algorithm[28] has been extended to include statistical analysis of proteome uniqueness as a function of mass accuracy and proteome size to evaluate the significance of search results (false identification rate).[29]

In this paper, we further expand this bioinformatics approach by analyzing spectra obtained by MALDI-TOF mass spectrometry from intact *Helicobacter pylori* (strain 26695). *H. pylori*, a Gram-negative bacterium, is found in stomach mucosa in a highly acidic environment, and is implicated as the causative agent of gastric ulcers and cancer. It is estimated that one-half of the world's population is infected, making this pathogen the most common bacterial infection. *H. pylori*'s genome was one of the first to be completely sequenced.[30] Direct sequence comparison of the genomes of two *H. pylori* strains was also performed after the sequence of a different strain (J99) had become available,[31] thus providing estimates on intraspecies genome plasticity. Very recently, protein–protein interaction maps of *H. pylori* have been generated from the complete genome sequence with specially developed bioinformatics tools.[32]

Several studies of *H. pylori*, employing mass spectrometry, have been published thus far.[33–38] Winkler et al. compared positive ion MALDI-TOF mass spectra from *H. pylori*, *Helicobacter mustela* and three *Campylobacter* strains.[33] These authors noted that the spectra, obtained from individual colonies, cultured in blood agar and subsequently suspended in 50% methanol–water, had unique biomarker patterns. Over 25 different ions from 2 to 62 kDa were observed, which permitted differentiation between the *Campylobacter* and *Helicobacter* species.[33] In an attempt to establish *H. pylori* strain-specific biomarkers in positive ion MALDI-TOF spectra, Nilson examined lysates from six different strains.[34] Owen et al. presented data on variations of MALDI spectra from intact *H. pylori* as a function of strain virulence.[35] In an initial proteomics study of *H. pylori* (strain 26695), McAtee et al. isolated by 2-D gel electrophoresis and characterized by mass spectrometry and genomic database search several 30 kDa proteins with the aim to identify potential vaccine candidates.[36] Nilsson et al. separated by rapid preparative electrophoretic procedures and characterized by MALDI-TOF more than 40 antigenic proteins with a typical M_r above 20 kDa from detergent-solubilized *H. pylori* extracts.[37] In an exhaustive proteome-wide study of three different *H. pylori* strains by high-resolution 2-D gel electrophoresis methodology, coupled to MALDI-TOF mass spectrometry, more than 100 of the most abundant proteins were identified and characterized.[38] Furthermore, 2-D electrophoretic patterns, incoporated in a dynamic 2D-PAGE image database and accessible over the Internet, give the option of interrogating each individual protein spot and providing on-line data for identified protein species[39]

There have been several objectives for the investigations reported here. First, we compile data on reproducible biomarker peaks, observed in MALDI-TOF from intact *H. pylori* (26695) samples. Cell lysis and extensive sample cleanup was typically the first step in most of the other MS studies reported until now. Second, by acquiring both positive and negative ion mode mass spectra, we demonstrate their utility for calibration and more accurate mass determination of biomarker peaks. Third, because the complete proteome of *H. pylori* (strain 26695) is available, it is used here as a model system to test Web-accessible algorithms for microorganism identification based on proteome database searches.[29,40] For instance, we illustrate the effects of proteome size and number of detected and matched biomarker peaks on the significance levels of microorganism identification. Furthermore, we examine the putative amino acid sequence of each biomarker peak, observed in spectra from intact *H. pylori*, and tentatively matched in the SwissPROT database.[41] In this manner, we evaluate approaches to account for posttranslational modifications (PTM), for example, N-terminal methionine cleavage, in order to improve identification reliability. We demonstrate that procedures to account for this PTM increase the significance of identification by an order-of-magnitude as a result of increasing the number of matched peaks.

EXPERIMENTAL SECTION

Microorganisms. *H. pylori*, strain 26695, was obtained from ATCC (Manassas, VA). The bacteria were grown in-house for 72 h using 2.5-L glass jars and tryptic soy broth medium with 10% horse serum (Sigma Chemical Co., St. Louis, MO). For generation of microaerobic growth conditions, "CampyGen CN25" paper sachets (Oxoid Ltd, Basingstoke, England) were placed in the jars. After harvesting, the material was purified by centrifugation at 10 000*g* for 10 min, and the pellet was washed with deionized water three times. The intact cells were lyophilized and stored at −20 °C prior to sample preparation and analysis. The experimental conditions for in-house growing of *E. coli*, strain 25404 (K-12), used as an external calibrant, have been already described.[27]

(26) Hathout, Y.; Ho, Y. P.; Ryzhov, V.; Demirev, P. A.; Fenselau, C. *J. Nat. Prod.* **2000**, *63*, 1492–1496.
(27) Ryzhov, V.; Fenselau, C. *Anal. Chem.* **2001**, *73*, 746–750.
(28) Demirev, P.; Ho, Y. P.; Ryzhov, V.; Fenselau, C. *Anal. Chem.* **1999**, *71*, 2732–2738.
(29) Pineda, F.; Lin, J.; Fenselau, C.; Demirev, P. *Anal. Chem.* **2000**, *72*, 3739–3745.
(30) Tomb, J., et al. *Nature* **1997**, *388*, 539–547.
(31) Alm, R. A., et al. *Nature* **1999**, *397*, 176–180
(32) Rain, J.-Chr., et al. *Nature* **2001**, *409*, 211–215.
(33) Winkler, M. A.; Uher, J.; Cepa. S. *Anal. Chem.* **1999**, *71*, 3416–3419.
(34) Nilsson, C. L. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 1067–1071.
(35) Owen, R. J.; Claydon, M. A.; Gibson, J.; Burke, B.; Ferrus, A. *GUT* **1999**, *45*, Suppl. 3, A28.
(36) McAtee, C.; Lim, M.; Fung, K.; Velligan, M.; Fry, K.; Chow, T.; Berg, D. *J. Chromatogr.* **1988**, *714*, 325–333.
(37) Nilsson, C.; Larsson, T.; Gustafsson, E.; Karlsson, K. A.; Davidsson, P. *Anal. Chem.* **2000**, *72*, 2148–2153.
(38) Jungblut, P. R.; Bumann, D.; Haas, G.; Zimny-Arndt, U.; Holland, P.; Lamer, S.; Siejak, F.; Aebischer, A.; Meyer, T. F. *Mol. Micorbiol.* **2000**, *36*, 710–725.
(39) http://www.mpiib-berlin.mpg.de/2D-PAGE/.
(40) http://infobacter.jhuapl.edu.
(41) http://www.expasy.ch.

CAUTION: *H. pylori* (26995) species is classified as a "biohazard level 2" (BL2) microorganism, and proper handling procedures should be followed.[42]

Sample Preparation. Samples were prepared according to standard procedures[28] as suspensions in acetonitrile/0.1% trifluoroacetic acid (TFA) (70/30, v/v) at a typical concentration of 5 mg/mL. Sinapinic acid (Aldrich Chemical Co., Milwaukee, WI) at 0.05 M was used as a matrix. Solutions of the matrix and sample (0.2 μL each) were mixed in individual wells of the stainless steel sample holder and allowed to dry prior to introduction into the interlock chamber of the TOF mass spectrometer. That corresponded to roughly 2×10^5 intact cells per sample deposited.[20] All of the sample preparation procedures involving *H. pylori* were performed in a laminar flow hood in a BL2-rated lab.

Mass Spectrometry. Both positive and negative ion mass spectra were obtained on a Kompact MALDI 4 (Kratos Analytical Instruments, Chestnut Ridge, NY) time-of-flight instrument in the linear mode at (±) 18 kV nominal accelerating voltage. Pulsed ion (delayed) extraction with a 0.3 μs delay time (optimized for ion focusing and transmission at m/z 10 000) was used for collecting spectra in both polarities. The fluence of the N_2 laser ("VSL-337ND", Laser Science Inc., MA, provided with the instrument) was ~10 mJ/cm^2 (4-ns pulse duration for 0.2 mJ energy/pulse at 337 nm laser wavelength). Each spectrum was a summation of 50 consecutive laser shots, with the beam rastered linearly across the sample surface. Internal (bovine insulin, bovine ubiquitin, equine cytochrome *c*) as well as external (*E. coli* K-12) mass calibration was used to provide mass accuracy better than 1 part in 2000. The proteins (Sigma Chemical Co., St. Louis, MO) were used as calibration standards without additional purification. Each individual protein was mixed with the bacterial sample/matrix solution on the sample holder in an amount sufficient to generate signals comparable in intensity with the microorganism biomarkers. Initial calibration (for both polarities) was performed by using only the protein standards or intact *E. coli* K12 cells (all calibration spectra obtained under identical instrumental conditions). A manual calibration procedure was developed for more accurate mass assignment of low intensity peaks (vide infra).

Database Search. A recently designed Web site[40] with interactive software for microorganism identification[29] has been accessed on-line. The software allows users to download subsets of a proteome database (e.g., the SwissPROT/TrEMBL database (release 38.0)[43]) containing bacterial proteins in a specified mass range. The partial proteomes of 18 microorganisms, those represented with at least 200 proteins in the range from 4 to 20 kDa, have been downloaded and used in the database search of experimentally obtained mass spectra (Table 1). Most of these microorganisms have completely sequenced genomes. Two *H. pylori* strains, 26695 and J99, are included in the set. However, the complete proteome of the J99 strain has not yet been translated; hence , the lower number of downloaded proteins, as compared with the 26695 strain. The mass tolerance used in the database search was ± 5 Da.

Table 1. Microorganisms Whose Partial Proteomes Were Downloaded and Used in the Web-Based Database Search[a]

microorganism[b]	genome size [Mb][c]	no. proteins (4–20 kDa) currently in SwissPROT
Mycosa pneumoniae	0.81	243
Chlamydia trachomatis	1.05	251
Rickettsia prowazekii	1.10	207
Treponema pallidum	1.14	251
Borrelia burgdorferi	1.44	470
Aquifex aeolicus	1.50	353
H. pylori J99[d]	1.64	291
H. pylori 26695	1.66	443
Thermotoga maritima	1.80	435
Haemophilus influenzae	1.83	492
Mycobacterium leprae[d]	2.80	656
Synechocystis sp.	3.57	911
B. subtilis	4.20	1420
Mycobacterium tuberculosis	4.40	1058
Salmonella typhimurium[e]	4.50	258
E. coli	4.60	2030
Pseudomonas aeruginosa[d]	6.30	200
Streptomyces coelicolor[e]	8.00	567

[a] Ref 40 [b] This library was compiled by requiring more than 200 protein entries for each organism in the mass range from 4 to 20 kDa. [c] Data compiled from TIGR microbial database – ref 45. [d] Proteome not completely translated. [e] Genome not completely sequenced.

RESULTS AND DISCUSSION

Protein Biomarker Spectra and Biomarker Mass Assignment. Positive and negative ion spectra of *H. pylori* (26695) are shown in Figure 1. Most of the detected peaks (35) are in the mass range below 20 kDa, although a relatively intense peak at ~26.4 kDa is discerned in spectra of both polarities. That peak is attributed to the urease α-subunit protein, one of the major *H. pylori* protein constituents. Its presence in MALDI-TOF spectra of protein extracts from *H. pylori* has been already noted.[37] There are a number of advantages in acquiring mass spectra in both polarities from the same sample and subsequently comparing the positive and negative ion mode spectra. For instance, in positive ion MALDI mass spectra of protein mixtures, one can easily distinguish between singly charged and multiply charged protein ions, since observation of multiply charged protein anions is much less likely. Using both positive and negative ion data allowed us to demonstrate that some unassigned peaks in already published positive ion spectra from *Bacillus subtilis* and *E. coli* correspond to the doubly protonated ions of molecular species present (e.g., peaks at m/z 4948 and at m/z 4775, 5149, and 5335 in Figures 1 and 2 of ref 28, respectively). Comparing the *H. pylori* spectra in Figure 1, we therefore conclude that most peaks correspond to singly charged individual biomarkers. The fewer number of peaks above m/z 10 000 in the negative ion mode is attributed to the lower sensitivity of the MALDI-TOF instrument for negative ions (due to, e.g., lower overall kinetic energy immediately prior to detection). In addition, in positive mode, a protein can form both protonated and sodiated molecular ion species. Their occurrence can be confirmed by the presence of a characteristic doublet having a 22 Da mass difference. In contrast, in the negative ion mode, the corresponding molecular ion will most often be a single peak.

(42) Centers for Disease Control and Prevention/National Institutes of Health. *Biosafety in Microbiological and Biomedical Laboratories*; 4th ed.; U.S. Government Printing Office: Washington, DC, 1999.
(43) Bairoch, A.; Apweiler, R. *Nucleic Acids Res.* **2000**, *28*, 45–48.

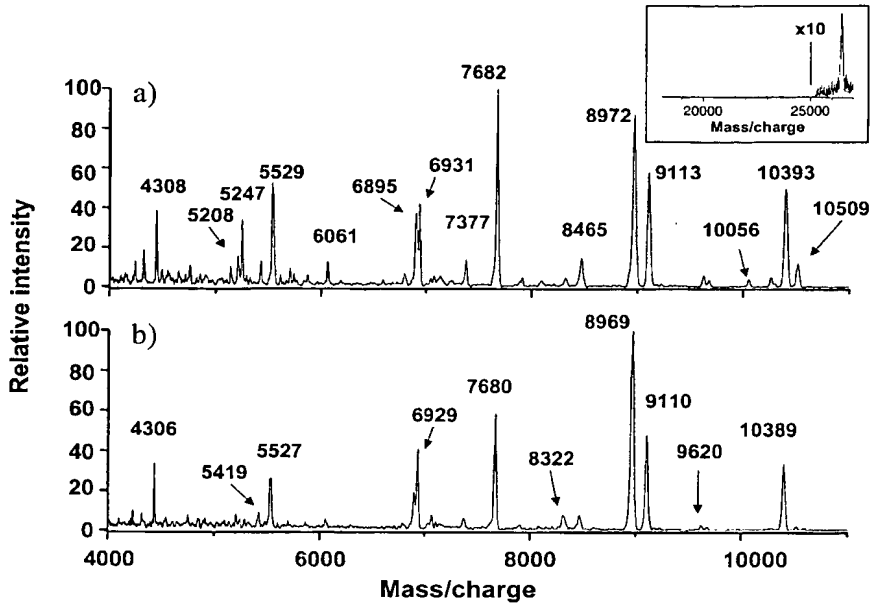

Figure 1. MALDI-TOF mass spectra from intact *H. pylori* in the *m/z* range 4000 to 11 000 in (a) positive and (b) negative ion modes. Inset: extended *m/z* range with the peak attributed to 26.4 kDa urease α-subunit protein.

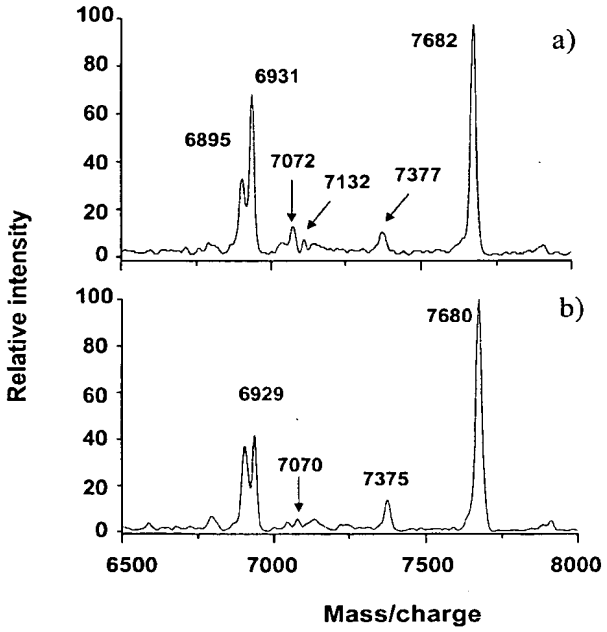

Figure 2. Expansion of the mass spectra in the *m/z* range from 6500 to 8000 in: (a) positive and (b) negative ion modes.

The calibration algorithm provided with the commercial system has been complemented by manual recalibration in order to improve the mass accuracy assignment for low-intensity peaks. Furthermore, since the pulsed ion extraction delay time is preset and constant, the use of calibration peaks separated by more than 2 kDa lowers the mass accuracy.[44] To "refine" the mass calibration for a protein mixture in a broader mass range, we split that range into smaller segments, typically within 2 kDa (Figure 2). We select sets of two calibration peaks (doublets) for more accurate calibration in the narrower range segments (between the doublets). These segments may overlap and cover the broader range between 4000 and 15 000. Peaks in the regions are used as controls. The sets of calibration doublets are chosen from intense peaks (e.g., at *m/z* 6931 and 7682, 7682 and 8972, 8972 and 10393; Figure 1a) that are close in mass in both the positive and negative ion spectra upon initial calibration. With such a stepwise calibration procedure and by averaging masses from spectra in both polarities, the masses of more than 30 individual biomarkers can be assigned with an accuracy better than ± 5 Da (Table 2). It is estimated (based on spectra of a mixture of protein standards) that the mass assignment of less intense peaks is improved by a factor of 2 with that procedure.

Partial Proteome Comparisons. The *H. pylori* genome is 1.66 Mb in size. The relative difference in microorganism genome sizes is reflected in their respective proteomes (Table 1). For instance, the number of potentially expressible proteins in the range between 4 and 20 kDa for *H. pylori* (strain 26695) is around 450, compared to more than 2000 proteins for *E. coli* in the same mass range. A comparison in the distribution of the 4 to 20 kDa range proteins for these two microorganisms is presented in Figure 3. For both microorganisms, the proteome densities[29] are quasi-uniform in that range, supporting the assumption in the theoretical model derivation.[29] Pairwise comparison between two proteomes can be performed by counting the number of proteins from each microorganism with M_r that overlaps with M_r of a protein from the other organism within a specified mass accuracy window. For these two organisms in the mass range from 4 to 20 kDa, more than 99% of the *H. pylori* proteins have unique M_r at 1 ppm accuracy, and only about 70% will differ in mass from the

(44) Kovtoun, S. V.; Cotter, R. J. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 841–853.

Table 2. Tentative Assignment of *H. pylori* Biomarker Peaks Based on Both Positive and Negative Ion Spectra

observed mass[a]	protein in SwissPROT	mass, Da	description	N-terminal amino acids	pI	remark
4306	P56058	4307	ribosomal	Met-Lys	11.09	b
5207						
5246	P56056	5246	ribosomal	Met-Lys	12.49	b
5420	O25662	5425	hypothetical	Met-Lys	9.70	b
5528						
5540						
5694						
5731						
5867						
6060	O25451	6058	hypothetical	Met-Lys	3.30	b
6894						
6930	P56051	6929	ribosomal	Met-Ala	12.21	c
7071						
7131						
7376						
7681	P56052	7683	ribosomal	Met-Lys	9.70	b
7919						
8098						
8218	P55974	8217	transl. initiator	Ala-Arg	9.46	b
8323	P94821	8318	hypothetical	Met-Ser	4.38	c
8464						
8971	Q9Z5L4	8975	cytotoxin assoc.	Val-Gly	5.30	b
9112	O25449	9113	hypothetical	Met-Asn	7.97	b
9230						
9623						
9676						
10055						
10255						
10391	P56022	10390	ribosomal	Met-Ala	10.00	c
10508						
11736						
13283	O26052	13287	hypothetical	Met-Lys	6.42	b
13467						
14034	P56018	14029	ribosomal	Met-Ala	10.33	c
14541	O25448	14542	flagellar	Met-Gln	5.03	c

[a] The neutral mass is listed. [b] Conforms with the PTM rules (see Scheme 1). [c] Does not conform with the PTM rules (see Scheme 1).

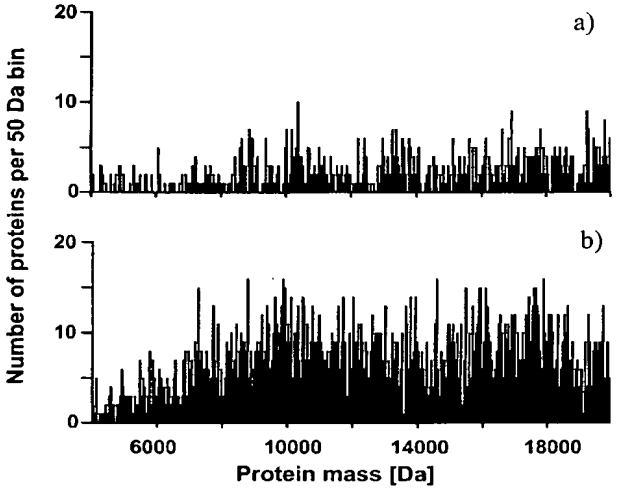

Figure 3. Comparison between protein M_r distributions for the partial (4–20 kDa) proteomes of *E.coli* (strain K12) and *H. pylori* (strain 26695); number of proteins/50 Da mass bins are plotted on the same vertical scale.

E.coli proteins at 100 ppm. This result underscores again the importance of accurate mass assignment of experimental data, as well as the need to include statistical criteria (e.g., significance testing[29]) in database search algorithms based on M_r. We also note the possibility for pairwise sequence comparison between entire proteomes of two individual microorganisms, by software available on The Institute for Genomic Research Web-site ("Genome versus genome protein hits"[45]). At 80% sequence similarity cutoff, only around 15 sequences in the entire proteomes of these two microorganisms can be matched.

Database Search. Using software available on the Web site,[40] initial search with the masses of the 35 biomarkers was performed. The "unknown" *H. pylori* 26695 was identified at a significance level better than 0.036 (Table 3). The significance level (ranging from 0 to 1) is a means to quantify statistically the probability for a random "hit" (i.e., experimental M_r overlapping a protein M_r from unrelated microorganism). It is a function both of proteome density and mass accuracy.[29] Its importance for reliable microorganism identification is well-illustrated with the current example. Although the number of hits for *E. coli* is larger than for *H. pylori* (18 versus 14), the fact that the latter has a less dense proteome is reflected in the much lower significance level, 0.036, and ultimately the correct identification. A significance level of 0.998 for *E. coli* means that all 18 peaks are "matched" by chance (due to the much higher *E. coli* proteome density, Figure 3). Another *H. pylori* strain, J99, is the runner-up with 10 matches and at 0.065 significance level (Table 3). Testing approaches for strain-specific microorganism identification, based on proteome database searches,

(45) http://www.tigr.org.

Table 3. Web-Based Identification Using a Total of 35 Biomarker Masses from the "Unknown" (*H. pylori* 26995)[a]

rank	organism	partial (4–20 kDa) proteome size	no. matches	significance level
1	*H. pylori* 26995	443	14	0.036
2	*H. pylori* J99	291	10	0.065
3	*M. leprae*	656	15	0.198
4	*R. prowazekii*	207	6	0.268
5	*H. influenzae*	492	11	0.348
6	*Th. maritima*	435	9	0.497
7	*Thr. pallidum*	251	4	0.788
8	*B. subtilis*	1420	19	0.818
9	*Synechocystis sp.*	911	12	0.919
10	*B. burgdorferi*	470	6	0.925
11	*Ps. aeruginosa*	199	2	0.935
12	*Str. coelicolor*	567	7	0.944
13	*M. pneumoniae*	243	2	0.971
14	*S. typhimurium*	258	2	0.978
15	*M. tuberculosis*	1058	11	0.990
16	*Chl. trachomatis*	251	1	0.997
17	*E. coli*	2030	18	0.998
18	*A. aeolicus*	353	1	0.999

[a] Ranked by significance level matching; post-translational modifications are not considered.

are beyond the scope of the present work. We also compare the pI's of the tentatively assigned protein biomarkers, most of which are basic (Table 2). However, observation of such species in both positive and negative ion mode spectra suggests that protein basicity is not a major factor, determining the observed biomarker spectral pattern. On the other hand, there are peaks in the spectra that are not matched by the *H. pylori* (26695) proteome. Several factors can be considered, including inaccurate mass assignment, posttranslational modifications, missequenced proteins, and proteins that are not present in the database. As already pointed out,[28] complementary information, including MS/MS data, can further facilitate microorganism identification.

Effect of Posttranslational Modifications on Microorganism Identification. Ribosomal protein synthesis in prokaryota starts with an N-formylated Met residue. Following the addition of several amino acid residues, the formyl group is almost invariably removed by the enzyme peptide deformylase.[46] The next processing step of the nascent polypeptide chain is cleavage of the N-terminal initiation Met amino acid. *N*-Met removal is the most common PTM for prokaryota, and it is estimated that ~50% of *E. coli* proteins undergo this specific PTM.[47] The activity of the responsible N-terminal bacterial aminopeptidases depends strongly on the N-terminal amino acid sequence.[48] In particular, the rates of Met-cleavage in *E. coli* have been correlated with the "penultimate" amino acid type.[46,48] Thus, the occurrence of this specific PTM can be cast into a set of empirical rules (Scheme 1). Correlations between biochemical processes involving cellular proteins and their amino terminal sequences have been reported previously. One such correlation is the "N-end rule" that maps bacterial protein half-life in vivo to the N-terminal amino acid.[49]

Scheme 1. Bacterial Aminopeptidase Cleavage Rules for N-terminal Met as a Function of the Penultimate (Xxx) Amino Acid Type (adapted from ref 48)

Post-translational N-terminal Met proteolysis:

$$NH_2 - Met \left| Xxx - \longrightarrow Met + NH_2 - Xxx - \right.$$

always cleaves if Xxx: Ala, Gly, Pro, Ser, Thr
loses activity if Xxx: Arg, Asn, Ile, Leu, Lys, Phe
variable activity if Xxx: Cys, His, Met, Trp, Tyr, Asp, Glu, Gln, Val

Posttranslational modifications (e.g., N-terminal Met cleavage) are not always reflected in proteome databases obtained from translation of the DNA open reading frames. For instance, there are 52 proteins listed in SwissProt as belonging to the ribosomal subunits of *H. pylori* (for both 26695 and J99 strains). All of these proteins are derived from gene sequences, and all contain N-terminal Met in their sequence. In contrast, the ribosomal proteins from *E. coli.* have been studied directly,[50] and losses of N-terminal Met from *E. coli* ribosomal proteins are already included in SwissPROT. For *E. coli*, the correct protein sequences and the correct M_r are listed in that database. If N-terminal Met is present in the database protein sequence, but is actually lost in a live organism, a mass difference of 131 Da will exist between M_r determined from the database and what is observed in the experimental mass spectrum. By examining protein sequences in the database, the fidelity of the M_r matching can be evaluated. If the sequence indicates that N-terminal Met should be retained (Scheme 1), then the mass match is considered significant. If not, the empirical mass is increased by 131 Da, and another search in the database is performed.

The sequences of the 52 ribosomal *H. pylori* proteins and the *N*-Met loss rules predict that one-half (26) should have undergone *N*-Met cleavage, and only 18 proteins should have *N*-Met intact. It also follows that several of the putatively identified proteins from Table 2 (those with experimentally determined M_r at 6930, 8323, 10391, 14034, and 14541) are predicted to have "lost" *N*-Met and, therefore, the calculated M_r will be reduced by 131 Da. For instance, the N-terminal amino acids in the database sequence of the protein P56022, tentatively assigned as the biomarker at mass 10391, are Met-Ala. According to the correlation (Scheme 1), the N-terminal Met should have been cleaved, suggesting that the initial match is not correct. The effect of this modification can be accounted for in an iterative procedure (Figure 4). For instance, the biomarker discussed above would correspond to a database protein with a mass of 10522 Da (increased by 131 Da). Database interrogation suggests a M_r match at 10522 ± 5 Da with a different protein, O24902. This is a plausible identification, since O24902 has a cleavable N-terminal Met (the database sequence starts with Met-Ser). Iteration results, applied to the experimentally observed *H. pylori* biomarkers from Table 2, are illustrated in Table 4. Consequently, and in order to extract significance level values, another on-line microorganism database search was performed with the "modified" list of plausible biomarker masses. The results

(46) Solbiati, J.; Chapman-Smith, A.; Miller, J.; Miller, Ch.; Cronan, J., Jr. *J. Mol. Biol.* **1999**, *290*, 607–614.
(47) Hirel, P. H.; Schmitter, J. M.; Dessen, P.; Fayat, G.; Blanquet, S. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 8247–8251.
(48) Gonzales, T.; Baudouy, J. *FEMS Microbiol. Rev.* **1996**, *18*, 319–334.
(49) Tobias, J. W.; Shrader, T.; Rocap, G.; Varshavsky, A. *Science* **1991**, *254*, 1374–1377.
(50) Arnold, R. J.; Reilly, J. P. *Anal. Biochem.* **1999**, *269*, 105–112.

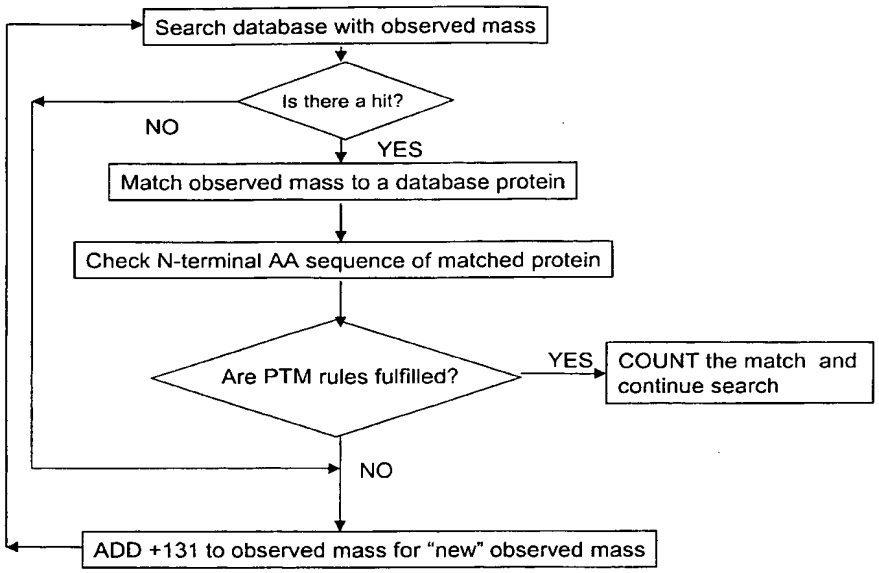

Figure 4. Tentative flowchart for including N-terminal Met cleavage in a proteome database search algorithm.

Table 4. Tentative Assignment of *H. pylori* Biomarker Peaks after Modifying the Observed Masses[a]

observed mass[b]	protein in SwissPROT	mass, Da	description	N-terminal amino acids	pI	remark
4306						
5338 (5207 + 131)	O25198	5335	hypothetical	Met-Ser	6.25	
5246						
5420						
5659 (5528 + 131)	P56054	5660	ribosomal	Met-Ala	10.86	c
5671 (5540 + 131)	Q48270	5669	hypothetical	Met-Glu	7.94	c
5825 (5694 + 131)						
5862 (5731 + 131)						
5998 (5867 + 131)						
6060						
7025 (6894 + 131)						
7061 (6930 + 131)						
7202 (7071 + 131)						
7262 (7131 + 131)	P56057	7260	ribosomal	Met-Pro	12.21	c
7507 (7376 + 131)	O25581	7512	oxalocrotonate	Met-Pro	6.03	c
7681						
8050 (7919 + 131)						
8229 (8098 + 131)						
8218						
8454 (8323 + 131)						
8595 (8464 + 131)	P56464	8590	acyl carrier	Met-Ala	.85	c
8971						
9112						
9361 (9230 + 131)						
9754 (9623 + 131)						
9807 (9676 + 131)						
10186 (10055 + 131)	Q9 × 5H7	10190	HELA	Met-Glu	4.34	c
10386 (10255 + 131)	O25689	10381	hypothetical	Met-Met-Glu	10.08	c
10522 (10391 + 131)	O24902	10517	hypothetical	Met-Ser	4.84	c
10639 (10508 + 131)						
11867 (11736 + 131)	P94838	11865	CAGC	Met-Lys	9.42	d
13283						
13598 (13467 + 131)	O25269	13596	CAG pathogen.	Met-Lys	9.95	d
14165 (14034 + 131)						
14672 (14541 + 131)						

[a] Including a putative posttranslational modification, N-terminal Met cleavage. See text for details. [b] The neutral mass is listed; 131 is added to biomarker masses listed in Table 2 that are not matched or do not conform to the PTM rules. [c] Conforms with the PTM rules (see Scheme 1). [d] Does not conform with the PTM rules (see Scheme 1).

are presented in Table 5. It is clear that accommodating this widespread posttranslational modification can increase the identification reliability for bacteria by at least an order of magnitude, because of a higher number of accurately matched peaks. Other

Table 5. Web-Based Identification Using Total of 35 Biomarker Masses from the "Unknown" (*H. pylori* 26995)[a]

rank	organism	partial (4−20 kDa) proteome size	no. matches	significance level
1	*H. pylori* 26995	443	17	0.002
2	*R. prowazekii*	207	6	0.268
3	*Thr. pallidum*	251	6	0.427
4	*B. burgdorferi*	470	10	0.434
5	*H. pylori* J99	291	6	0.567
6	*S. typhimurium*	258	5	0.638
7	*B. subtilis*	1420	20	0.717
8	*H. influenzae*	492	8	0.774
9	*Chl. trachomatis*	251	4	0.786
10	*A. aeolicus*	353	5	0.867
11	*E. coli*	2030	23	0.893
12	*M. pneumoniae*	243	3	0.899
13	*Synechocystis sp.*	911	12	0.919
14	*Ps. aeruginosa*	199	2	0.935
15	*Th. maritima*	435	5	0.952
16	*M. leprae*	656	7	0.980
17	*M. tubreculosis*	1058	11	0.990
18	*Str. coelicolor*	567	5	0.992

[a] Ranked by significance level matching; loss of N-Met is considered. See text for details.

less common PTM could be also considered by this identification strategy, provided that biochemical rules correlating the PTM with, for example, the protein sequence, are available.

CONCLUSIONS

MALDI-TOF spectra from intact *H. pylori* species contain sufficiently high numbers of biomarker peaks to allow the correct microorganism identification by Internet-accessible proteome database search algorithms. Acquiring mass spectra in both polarities from the same sample results in more accurate biomarker mass assignment and improves the overall reliability of the method. The importance of advanced classification criteria for the assessment of search results is experimentally illustrated. It is confirmed that statistical significance testing, introduced earlier, reduces the possibility of false identification for microorganisms with less dense proteomes. Furthermore, we propose a procedure to account for additional data contained in genome-derived proteome databases. The N-terminal amino acid sequences of putatively identified proteins are correlated with N-terminal Met removal using empirically established rules. The sequence signals for posttranslational enzymatic cleavage of N-terminal Met, the most common PTM for prokaryota, are iteratively incorporated in the database search to evaluate the effect on microorganism identification. It is demonstrated, on the basis of the example studied here, that the reliability of microorganism identification is improved by at least an order of magnitude. We also note that an alternative algorithmic approach is to modify the proteome database in a fashion that takes into account the frequency of this particular posttranslational modification, quantifying the empirically established rules.

ACKNOWLEDGMENT

We thank Amy Freas (University of Maryland) for growing the microorganisms used in this study. P.D. and C.F. acknowledge financial support from DARPA. F.P. and J.L. performed this work under U.S. Navy contract N00024-98-D8124.

Received for review April 24, 2001. Accepted July 24, 2001.

AC010466F